HumRRO

Special Report

HumRRO
SR-WATSD-99-39

June/July, 1999

Documentation of Third Party Checking of 1998 Pre-equating for Kentucky Core Content Test: IRT Scaling of Multiple Choice and Open Response Test Items

R. Gene Hoffman
Arthur A. Thacker
James R. McBride

Prepared for:

The Kentucky Department of Education
500 Mero Street
Frankfort, KY 40601

Contract Number BP 010492

Human Resources Research Organization
66 Canal Center Plaza, Suite 400 • Alexandria, VA 22314-1591

Documentation of Third Party Checking of 1998 Pre-equating for Kentucky Core Content Test: IRT Scaling of Multiple Choice and Open Response Test Items

Table of Contents

Documentation of Third Party Checking of 1998 Pre-equating for Kentucky Core Content Test: IRT Scaling of Multiple Choice and Open Response Test Items

Introduction

In order to make the transition from the now-defunct KIRIS test to the Kentucky Core Content Test with the minimum amount of disruption a system of linking the old test with new was necessarily devised. This link allows Kentucky to maintain the consistency of its student performance levels and to apply the student Kentucky Core Content Test scores to a newly revised accountability calculation. The main difficulty in linking the two tests is that KIRIS only applied student scores on the open-response section of the test towards a school's accountability index and toward individual student performance levels. The Kentucky Core Content Test will use both open-response and multiple-choice format questions to make those determinations. Students will still receive ratings in terms of the Novice, Apprentice, Proficient, and Distinguished levels of performance, but multiple-choice questions will now be included in those determinations.

Purpose and Organization of the Report

The purpose of the report is to provide documentation of HumRRO's activities for checking the scaling multiple-choice and open-response items from the 1998 Kentucky Instructional Results Information System (KIRIS) and equating the results to open-response only scores from KIRIS. For Reading, Mathematics, Science, and Social Studies, the process involved:

1. Creating working files (PARDUX *.rwo) that combine 1998 multiple-choice and open-response data for students in the 1998 open-response calibration sample.

2. Preparing control files (PARDUX *.ctl), and then using CTB's PARDUX program, scaling multiple-choice and open-response items together and scoring students using only their open-response items.

3. Transforming students' 1998 original open-response adjusted (1993 metric) thetas to a 500/50 scale and creating a frequency distribution.

4. Equating scores from Step 2 to scores from Step 3 using the linear approximation to equipercentile routine in CTB's FLUX program.

5. Confirming the equating constants from Step 4 match those derived by CTB.

For reading, mathematics, science, social studies, the following documentation is provided:

A. Flux Log—(includes plots (M1 & M2) and percentile equivalence tables). The Flux Log contains the equipercentile plot that might best be considered the final results of this procedure. The M1 and M2 values printed on the graphs respectively represent the slope

and Y-intercept for the equipercentile line. The plot itself gives a good indication of the fit between the previously calculated KIRIS scores and parameters of those same items placed on a scale that also contains the multiple choice questions. Another indication of the fit of the data is the combined cumulative frequency distribution graph, also presented in this section. The previous Kentucky data is plotted along with the transformed data. If the data were to match perfectly, one line would be atop the other. For most tests, the data match very closely. Also included is the actual output of the Flux program for the linking. This section includes the scale scores that match percentile ranking for both the transformed data set and the previous Kentucky data.

B. Pardux Log of Open Response Only Theta Estimation. The Pardux Log of Open Response Only Theta Estimation is the output from Pardux for the estimation of thetas using only open response items. It includes the file name from which the item parameters were read and pertinent item counts. It also contains the filename for the new control file used in this procedure.

C. Pardux Log of Item Parameter Estimation. The Pardux Log of Item Parameter Estimation is the best initial check to ensure that all items were correctly read into the program. It contains an overall mean and standard deviation as well as the stage at which convergence occurred. The file also contains fit statistics regarding the parameter estimation.

D. Item Parameters Text File. The Item Parameters Text File contains information about specific items by item number. This file serves as a quick documentation of items that were problematic during the parameter estimation. Multiple choice item parameters are supplied in 3PL metric. Open response items use the 2PPC model.

E. Control file (*.ctl) printout showing the identification of *.rwo items as multiple-choice, open-response, or unused (pre-test or item previously identified by KDE as problematic, e.g., negative biserial).

F. Notes Regarding Problem Items. Also included in the printed documentation are notes regarding problem items. When item parameters are estimated, there are occasionally items that do not fit the model exactly. Some of these items exhibit a very high alpha parameter. Others do not fit the function used in the estimation very well and are not estimated. These problem items are examined and either left as is, or some form of additional manipulation is required. In extreme cases, items must be "hand-fit" in order to obtain estimates from them. In these cases, we must rely on our notes to document what was done to these items in order to include them in the estimation and to allow for additional checking in the future.

For Art&Humanities and Practical Living/Vocational Studies, no equating procedures were employed. The cut points for the Kentucky Core Content Test will be determined by arranging the scores to approximate the proportions of students in each scoring category from the 1998 KIRIS test. KDE/Robert Wetter, in email dated June 23, 1999, provided frequencies. See Table 1 below.

Table 1 NAPD item score distributions from 1998						
	Arts & Humanities			Practical Living/Vocational Studies		
	Grade 5	Grade 7	Grade 11	Grade 5	Grade 7	Grade 11
Novice	68%	53%	52%	38%	63%	54%
Apprentice	29%	40%	44%	56%	30%	39%
Proficient	1%	4%	2%	5%	5%	4%
Distinguished	2%	2%	2%	1%	2%	2%

No further documentation for Arts&Humanities and Practical Living/Vocational Studies was provided.

In addition to this report, electronic copies of additional, supplementary output from PARDUX and FLUX runs have been provided to Kentucky Department of Education. Due the size of these files, they are only supplied electronically. The electronic files include the printed logs and the following files.

1. RWO File
2. Multiple-Choice + Open-Response PAR File
3. Multiple-Choice + Open-Response Status Text File
4. Multiple-Choice + Open-Response Distribution Text File
5. Multiple-Choice + Open-Response Parameters Text File
6. Multiple-Choice + Open-Response SE's File
7. Multiple-Choice + Open-Response FitQ1 File
8. Multiple-Choice + Open-Response FitQQ3 File
9. Multiple-Choice + Open-Response Theta Estimates in VEC File
10. Multiple-Choice + Open-Response Theta Estimates in TST File
11. Multiple-Choice + Open-Response Parameter Estimation Summary File
12. Multiple-Choice + Open-Response Parameter Estimation Details File
13. Open Response Only Control File
14. Open Response Only Theta Estimates in TST File
15. Open Response Only Theta Estimates in VEC File
16. Frequency (FRQ) File
17. SAS Program for Creating FRQ File From RWO File
18. HLK File (Used for multiple grades and subjects)

All files are identified by subject area and grade and placed in separate electronic folders. They are provided in .ZIP format for convenience. For a more complete description of these files see Appendix A and the documentation provided with Pardux and Flux.

Results

After performing periodic checks with CTB as individual tests were scaled, HumRRO and CTB reached agreement on the scaling constants of Reading, Math, Science, and Social Studies. In fact, the equating constants (final results of the process) computed by CTB and HumRRO match exactly for all but 5th grade social studies. This difference is not large enough to be considered significant, however, we have investigated the reason for this discrepancy and discovered an idiosyncrasy in the Pardux computer program.

Table 2 presents the percentile ranges chosen by CTB and the equating constants reached through the scaling process. (The actual e-mailed document from CTB that included this table is included in the Appendix.) Table 3 contains the same information as determined by HumRRO. HumRRO did not select the percentile ranges independently, but instead used those chosen by CTB.

Table 2. Percentile Ranges and Equating Constants Computed by CTB

Subject	Grade	Percentile Range	M1	M2
Mathematics	5	7 through 90	.659	221.919
	8	10 through 95	.627	216.911
	11	14 through 99	.705	179.006
Reading	4	1 through 99	.629	233.873
	7	1 through 99	.636	196.415
	11	1 through 99	.653	232.634
Science	4	11 through 99	.554	264.041
	7	1 through 99	.523	239.112
	11	6 through 99	.550	267.383*
Social Studies	5	1 through 99	.639	216.904
	8	11 through 99	.727	146.174
	11	1 through 99	.839	129.120

* Note. This Table was provided prior to discovering an error in 11th Grade Science. CTB's new M2 = 267.305.

Table 3. Equating Constants Computed by HumRRO

Subject	Grade	M1	M2
Mathematics	5	.659	221.919
	8	.627	216.911
	11	.705	179.006
Reading	4	.629	233.873
	7	.636	196.415
	11	.653	232.634
Science	4	.554	264.041
	7	.523	239.112
	11	.550	267.305
Social Studies	5	.638*	217.213*
	8	.727	146.174
	11	.839	129.120

* Note. These results do not match CTB due to an idiosyncrasy of the Pardux program. The results match CTB if the idiosyncrasy is accounted for in the procedure. These results are included in the report to indicate the inconsistency and to remind us that the difference should be more thoroughly investigated.